Computer Vision and Image Understanding 113 (2009) 353–371

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Unsupervised view and rate invariant clustering of video sequences ☆

Pavan Turaga *, Ashok Veeraraghavan, Rama Chellappa

Department of Electrical and Computer Engineering, Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 30 October 2007
Accepted 18 August 2008
Available online 11 September 2008

Keywords:
Video clustering
Summarization
Surveillance
Cascade of linear dynamical systems
View invariance
Affine invariance
Rate invariance

ABSTRACT

Videos play an ever increasing role in our everyday lives with applications ranging from news, entertainment, scientific research, security and surveillance. Coupled with the fact that cameras and storage media are becoming less expensive, it has resulted in people producing more video content than ever before. This necessitates the development of efficient indexing and retrieval algorithms for video data. Most state-of-the-art techniques index videos according to the global content in the scene such as color, texture, brightness, etc. In this paper, we discuss the problem of activity-based indexing of videos. To address the problem, first we describe activities as a cascade of dynamical systems which significantly enhances the expressive power of the model while retaining many of the computational advantages of using dynamical models. Second, we also derive methods to incorporate view and rate-invariance into these models so that similar actions are clustered together irrespective of the viewpoint or the rate of execution of the activity. We also derive algorithms to learn the model parameters from a video stream and demonstrate how a single video sequence may be clustered into different clusters where each cluster represents an activity. Experimental results for five different databases show that the clusters found by the algorithm correspond to semantically meaningful activities.

1. Introduction

Recent years have seen a tremendous explosion of multimedia content fueled by inexpensive video cameras and the growth of the internet. The amount of video data being recorded and accessed by the general populace has been increasing with the rising popularity of several video sharing websites. Several technical challenges need to solved to enable efficient search and retrieval in such large and often unstructured datasets. In the context of videos, most video-sharing websites feature user-supplied tags. It is assumed that the tags describe what is going on in a video. This has several drawbacks, since tags are subjective and the same meaning may be conveyed by a multitude of tags. Further, it is not uncommon to encounter tags whose meaning is not known to a user and are sometimes irrelevant to the video. The idea of user generated tags would also not scale with the increasing size of the dataset. Instead of providing such textual descriptions, we propose to analyze the patterns of motion in a video and automatically extract 'clusters' which when visually presented to a user can convey maximum information about the contents of the video. For example, given a tennis video, short segments depicting elements such as fore-hand, back-hand, smash, etc., when visually presented to a user would convey more information than a set of textual descriptions.

Unsupervised activity-based indexing goes far beyond the traditional problems of activity analysis and recognition, where one knows what one is looking for. Unsupervised indexing requires that activity patterns be discovered without deciding a priori what to look for. As a motivating example, consider the problem of understanding a foreign language. If one hears only a continuous stream of words, how does one know where a word begins and where it ends. If one knew the words, the boundaries between them can be easily perceived. And if one knew the boundaries, then the words can be learnt as well. Similarly, given a continuous video stream, if we knew what activities occur in it, we can discover the boundaries between them—and if we were given the boundaries, the individual activities could be learnt as well. In the context of activities each action primitive is composed of a coherent set of features, and an activity is defined by the way the primitives are put together. Activity-based indexing can benefit by gaining insight into how humans perceive and recognize activities. First, we discuss a general framework of activity perception. Then, we discuss how the cascade of linear dynamical systems model (CLDS) can be derived from the proposed framework.

Applications for automatic discovery of activity patterns are numerous. For example, security and surveillance videos typically have repetitive activities. If the typical activities can be clustered, then several problems such as unusual activity detection, efficient indexing and retrieval can be addressed. Forensic analysis of surveillance videos is another fast growing and important application

☆ Some parts of this work were presented at CVPR 2007 [1]. This research was funded (in part) by the U.S. Government VACE Program.

* Corresponding author.
E-mail addresses: pturaga@umiacs.umd.edu (P. Turaga), vashok@umiacs.umd.edu (A. Veeraraghavan), rama@umiacs.umd.edu (R. Chellappa).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.08.009

area. Current approaches to video forensics involve linear searches over the entire video feed by a human analyst and hence are not scalable when there are a large number of cameras deployed at various locations. Instead of expecting an analyst to sift through the voluminous data, we ask—can 'clusters' of activities be presented that embody the essential characteristics of the videos? The need for such activity-based indexing stands to increase in the near future as more security installations are deployed in a wider variety of locations.

Traditionally, the problem of recognizing human activities from video has received more attention than automatic discovery. One of the earliest experiments demonstrating the richness of spatio-temporal features was done by Johansson [2] who showed that it was possible to recognize humans based on their motion. Since then there has been an explosion of research in computer vision into automatic analysis and recognition of human activities from video. Parallel to the development of accurate and efficient recognition techniques, there has also been a lot of interest in automatic discovery of patterns from raw data. Pattern recognition vs pattern discovery is a fundamental choice that is faced in almost all areas of machine learning. Specific to the activity analysis area, existing literature focuses on the recognition problem to a large extent. In a largely unrelated setting, there has been significant research into indexing of multimedia data such as news clips, sports videos, etc., according to their content such as in [3]. The pattern discovery approach has also been pursued for this problem domain such as in [4]. In practical applications it is often a combination of supervised and unsupervised approaches that yields the best results [4]. In such cases, an 'unsupervised' approach would involve a very small amount of supervision. The approach we present is largely unsupervised where we assume very little about the data and the supervision is maintained at very low-levels, for example, in the choice of features, temporal segmentation criterion, etc.

1.1. Organization of the paper

The rest of the paper is organized as follows: first, we present a general framework of activity perception in Section 2 and draw connections to computational and mathematical models in Section 3. Then, we propose the CLDS model for the specific purpose of activity-based mining in Section 4. The problem of learning the model parameters is addressed in Section 5. Further, in Section 6 we show how invariances to view, affine transforms and execution rate can be built into the mining algorithm. Finally, in Section 7 we describe several experiments to demonstrate the effectiveness of our algorithms.

2. Perception of activities

In this section, we propose a general framework for activity perception and recognition, from which specific algorithms can be derived. The perception of activities can be seen as proceeding from a sequence of 2D images to a semantic description of the activity. Activity perception can be naturally decomposed into the following three stages:

(1) Dynamic sketch
(2) Action sketch, and
(3) Semantic sketch.

2.1. Dynamic sketches

The purpose of early stages of vision is to construct primitive descriptions of the action contents in the frame. These primitive descriptions must be rich enough to allow for inference and recognition of activities [5]. Most of the sensory information that is available in videos is actually uninteresting for the purpose of activity-based video indexing and only serves to confound the latter stages of the algorithms. One very important characteristic of this stage is to weed out all the unnecessary sensory information and retain just those elements that are relevant for activity-based video indexing. Visual encoding mechanisms present in the human brain mimic this phenomenon and this is called predictive coding. Barlow [6] and Srinivasan et.al. [7] contend that predictive coding is not just a mechanism for compression but actually goes much further and enables animals to process information in a timely manner. We refer the interested reader to early works of Barlow [6] on the importance of this stage of visual processing in order to enable vision systems to react and process information in a timely manner.

2.2. Action sketch

Studies into human behavior show that human actions can be temporally segmented into elementary units, where each unit consists of functionally related movement [8]. For example, a car parking activity may be considered to be formed of the following primitives—'Car enters a parking lot', 'Car stops in the parking slot', 'Person walks away from the car'. Such a description requires the ability to segment an activity into its constituents and then develop a description for each of the constituent actions. Each constituent action is like a word describing a short, consistent motion fragment. Hence, this stage can be interpreted as providing a 'vocabulary' with which to create sentences (activities). *In the remainder of the paper, by 'action' we refer to a short segment of consistent motion, whereas, by 'activity' we refer to a composition of such actions that leads to an activity.*

Representing activities using such linguistic models has been in existence in various other fields and disciplines. Several dance notation schemes are used in practice to interpret complex dance moves. Though not extremely detailed, they are easy to interpret and reproduce in actual steps. It has also been found that the most commonly observed human activities in surveillance settings such as reaching, striking, etc., are characterized by distinctive velocity profiles of the limbs that can be conveniently modeled as a specific sequence of individual segments—constant acceleration followed by constant velocity followed by constant deceleration [9]. This lends credence to the fact that human actions can be modeled as a sequence of primitive actions, where each action is governed by a simple model.

2.3. Semantic descriptions

Semantic descriptions perform the same function as grammatical rules for a language. They detail how several constituent action primitives may be combined together in order to construct or recover complex activities. The most common rules for creating complex activities from constituent actions are sequencing, co-occurrence and synchronization. For example, a single-threaded activity can be said to consist of a linear sequence of a few primitives. An example of a single-threaded activity is 'Person approaches a door' → 'Person swipes the access card' → 'Person enters a building'. Similarly, a complex multi-threaded activity can be seen as a collection of several single-threaded activities with some constraints such as concurrence and synchronization among them. Thus, this stage can be seen as providing the rules for combining the primitives—similar to a set of grammatical rules needed to construct meaningful sentences from individual words.

In the next section, we draw connections with computational approaches and show how several well-known mathematical tools can be used at each of these stages.